Slide 1 of 37

Title Slide: PIAAC Sample Design, Weights, Variance, and Missing Data

Slide 2 of 37

This module summarizes the sample design for the Program for the International Assessment of Adult Competencies (PIAAC) as it relates to study weights, and provides information about how to appropriately use weights and variance estimation procedures with PIAAC data. Additionally, the module will explain how scaling is used in the large-scale international assessments and how plausible values are used when analyzing assessment data.

PIAAC used a complex design for selecting the nationally representative sample of adults ages 16 to 65. As discussed in the common modules, when analyzing data from complex sample surveys, certain procedures must be used to assure that results are representative of the target population and that hypotheses tests are accurate. Specifically, weights must be applied, and standard errors must be computed in a way that takes the complex sample into account. Furthermore, plausible values must be used while analyzing the assessment portion of the data. This module discusses these topics specifically in relation to the analysis of PIAAC data.

Additionally, the module describes how missing data are handled in the data file to ensure accurate data analysis.

For a general review of sampling weights for NCES datasets and procedures for calculating standard errors, visit the common modules titled “Statistical Analysis of NCES Datasets Employing a Complex Sample Design” and “Analyzing NCES Complex Survey Data” by clicking on the corresponding underlined screen text ‘calculating appropriate standard errors’ and ‘weights that must be applied’, respectively.

Slide 3 of 37

The specific topics that will be covered in this module include:

- The PIAAC target population;
- The PIAAC sample design, including the frame, sampling units, sample size, and eligibility criteria;
- Quality standards of the PIAAC sample and assessment designs;
- An overview of the PIAAC assessment design and eligibility criteria;
- PIAAC sampling weights, including non-response and other adjustments, the final weight and the corresponding replicate weights;
- Proficiency estimation in PIAAC, including estimation of the statistic and variance using plausible values, or PVs; and,
- A brief discussion of how missing data are handled in the data file to ensure accurate data analysis.

Slide 4 of 37

As discussed in earlier modules, the PIAAC target population consists of all non-institutionalized adults, 16 to 65 years old, who reside in the country at the time of data collection, regardless of citizenship, nationality or language.

Residence is defined as living in a house, an apartment, or a condo as well as workers' quarters, halfway homes, and dormitories, fraternities, or sororities.

Non-institutionalized adults are defined as people who are not residing in prisons, hospitals, nursing homes, or military barracks and bases.

In addition to the minimum sample of 5,000 adults, countries were allowed to oversample certain portions of their populations. Australia, Canada, the Czech Republic, Denmark, Germany, and Poland oversampled various subpopulations, for example adults aged 16 to 24, those aged 55 to 65, or those living in certain geographical areas.

Countries were also allowed to include additional subpopulations of national interest that were not included in the original PIAAC sample. Australia and Denmark, for example, added a sample of older adults, aged 64 to 74, and a sample of persons that had participated in the 2000 administration of the Program for International Student Assessment, or PISA.

It is important to note that neither the subpopulation additions nor the oversampled portions of the populations are included in the international files.

Slide 5 of 37

The PIAAC standard sample design is a self-weighting design of persons. When every person in the population has an equal probability of selection, each sampled person carries the same weight, and is given the same sampling weight for statistical calculations. It is important to note that this does not mean that PIAAC uses a simple random sample design.

Participating countries were required to develop their sample design and selection plans according to the standards provided in the PIAAC Technical Standards and Guidelines to be approved by the Consortium. The Consortium allowed each country to choose the most optimal and cost effective sample design and selection approach, as long as the design applied full selection probability methods to select a representative sample from the PIAAC target population.

Geographically small countries chose sample designs with less clustering and fewer stages of sampling. For example, most of these countries employed stratified simple random sampling or stratified systematic sampling. For geographically large countries, the typical sample design was a multistage, stratified clustered area sample, employing selection with probability proportional to size.

Slide 6 of 37

PIAAC standards require that sampling frames be up to date and include only one record for each member of the target population. Several types of frames were used by participating countries to select the household samples. Some countries used a list of residents or a national population registry as the sampling frame, while others employed a master sample already employed for national surveys. Other countries used a stratified area sample design in which a frame of households was created within selected geographic clusters.

As the ultimate sampling unit, each person in the PIAAC target population must have a calculable, nonzero probability of selection; that is, he or she must have a chance of being selected into the PIAAC sample. The Organization for Economic Cooperation and Development, or OECD, sent recommended formulas to participating countries for selecting the ultimate sampling unit for one-, two-, three-, and four-stage sample designs before countries began the sample selection process. Countries were asked to either confirm their use of those formulas, or to provide alternate formulas showing their deviations from the self-weighting design. Ultimately, participating countries used either one-stage stratified or non-stratified sampling; or two-, three- or four-stage stratified, probability proportionate to size sampling designs.

Slide 7 of 37

Eight countries, including Austria and Sweden, implemented a one-stage sample design, in which there is only one sample unit… persons.

Slide 8 of 37

The sampling units for countries with two-, three-, and four-stage sample designs included entities such as districts, communities, households, municipalities and dwelling units. These sample designs are described in more detail in Chapter 14, Section 14.4 of the Technical Report of the Survey of Adult Skills. This report can be accessed by clicking on the corresponding underlined screen text.

Slide 9 of 37

The United States used a four-stage stratified area sample. The first stage consisted of primary sampling units, or PSUs, comprised of counties or groups of contiguous counties. The secondary sampling units, or SSUs (referred to as segments) consisted of census area blocks. The third stage included housing units, or DUs, containing households; and Stage 4 involved sampling eligible person(s) within households.

A screener that collected the age and gender of all household members was used to determine person-level eligibility.

Slide 10 of 37

The minimum sample size requirements for the standard target population speaking the main language of the country was dependent on which, if any, of the optional components of the assessments were administered in the country. If both the Problem solving in technology-rich environments (or PS-TRE) and reading components were administered; or, if only PS-TRE was administered, then a minimum of 5,000 completed cases was required.

If only reading components were administered; or, no optional components were administered, then a minimum of 4,500 completed cases was required. More information about the components of the PIAAC study can be found in the module titled, 'Data Collected Through the PIAAC' which can be accessed by clicking on the underlined screen text, 'components.'

Each country that tested in another language or languages, in addition to the main language, had to add completed cases proportional to the number of people speaking the additional languages in the country.

Countries planning to report national level proficiency, regardless of the languages tested, had to achieve the appropriate minimum completed sample size for their main language.

Thus, the minimum sample size requirement for an individual country not only depended on the optional assessments administered, and the number of languages being tested, but also on the number of reporting languages determined by the country.

Slide 11 of 37

Most countries conducted both the PS-TRE and reading components, in addition to the required literacy and numeracy assessments.

However, Finland, Japan, and the Russian Federation conducted the PS-TRE only; Cyprus, Italy and Spain conducted the reading components only; and France declined both optional assessments.

Five countries performed the assessment in multiple languages:

- Canada performed the assessment in Canadian English and in French, with about 10,000 completed cases;
- Estonia performed the assessment in Estonian and Russian, with about 7,500 completed cases;
- Finland performed the assessment in Finnish and Swedish, with about 5,280 completed cases;
- the Slovak Republic performed the assessment in Slovak and Hungarian, with about 5,550 completed cases; and

- Spain conducted the assessment in five languages – Castellano, Gallego, Catalan, Valencian, and Euskera – with about 6,000 completed cases.

Slide 12 of 37

The United States administered the required components of literacy and numeracy, and the optional PS-TRE and reading components, in English, as the only language for the assessment. Therefore, in Round 1 of the PIAAC data collection, the U.S. target sample size was 5,000 adults. To reach this target, 9,468 households in the U.S. were sampled and 6,916 of these were eligible for PIAAC. From these households, 4,898 adults completed the background questionnaire and 4,820 completed the assessment.

Slide 13 of 37

To be eligible for participation in PIAAC at the sampling stage, households had to be units with occupancy; that is, not vacant. In addition, they had to be intended for regular occupancy, and not just for seasonal use; actual dwelling units, as opposed to institutions; and, housing persons 16 to 65 years old. Units that did not meet these requirements were considered ineligible for the survey, and had to be accounted for in the derivation of the final sample size. In the U.S., the occupancy rate was 86 percent, and the eligibility rate was 82 percent.

The expected response rates reported during the National Survey Design and Planning Report process were taken into account to ensure that the initial sample sizes were large enough to yield the required number of assessments. The Consortium encouraged each country to consider selecting a reserve sample of 10% or more of the size of the main, original sample. The requirement was to select the reserve sample at the same time as the original sample, and then set it aside and not use it unless sample monitoring showed potential for shortfall. Reserve samples were recommended over supplemental samples because computing the selection probabilities would be simpler with a reserve sample than it was with supplemental samples. The same process was used if a country was concerned about exceeding the target sample size by a significant amount.

Slide 14 of 37

To ensure high quality and high comparability of the data across all participating countries, PIAAC developed a comprehensive set of quality assurance and quality control checks for all sampling activities, including sample design and selection results. Most countries adhered to the standards. Those that did not, received cautionary remarks on the sample selection process, either during the theoretical stage in the home office, or in the field. More information regarding the standards can be accessed by clicking on the underlined screen text, ‘standards.’

For example, cautionary remarks to the home office were given to the Czech Republic for late sample selection forms, to Germany for simulated probabilities of selection, and

to Japan for an approved deviation from the previously defined Technical Standards and Guidelines, given the disastrous earthquake.

Cautionary remarks to the home office AND for the in-field sampling were given to Australia based on unknown quality level due to country confidentiality restrictions or unavailability of data.

Cautionary remarks for the in-field sampling were also given to the United Kingdom for theoretical person base weights for 52 cases, 49 in England and three in Northern Ireland, that were imputed due to a technical problem with the contact data that the interviewers entered.

Slide 15 of 37

PIAAC also developed a comprehensive set of quality assurance and quality control checks for all assessment implementation procedures. All countries adhered to the standards, resulting in the general cross-country comparability of the assessments. For more information on the assessment design procedures refer to Chapter 1 of the PIAAC Technical Report, which can be accessed by clicking on the underlined screen text, ‘assessment.’

Slide 16 of 37

Although details about the PIAAC assessment design are covered in more detail in the module titled “Data Collected Through the PIAAC,” it is important to recap a few points about the assessment design here, as they relate to the discussion of PIAAC variance estimation that follows in this module.

The PIAAC cognitive, or psychometric, assessment items were administered to participants via either a computer-based assessment or a paper-based assessment. Multiple matrix sampling was used so that each sampled individual received only a subset of cognitive items. In addition, the PIAAC assessment used a multi-stage adaptive design, in which the sets of cognitive items that each respondent received were based on the respondent’s previous answers. PIAAC is the first international comparative survey to apply adaptive testing procedures.

Slide 17 of 37

PIAAC’s complex survey sample design and its complex assessment design require multiple weights to obtain valid survey estimates, and plausible values to obtain valid assessment estimates. Validity, in this case, means representative of the target population.

The complex assessment design is described in detail in the module titled “Data Collected Through the PIAAC.” You can access this module by clicking on the corresponding underlined screen text.

Let’s review the purpose of sampling weights.

Slide 18 of 37

Sampling weights are designed to permit unbiased estimates and make the data representative of the target population by:

- Compensating for the possible disproportionate sampling of various subgroups in the population;
- Compensating for non-coverage in the sample, due to inadequacies in the sampling frame or other reasons for non-coverage;
- And, reducing sampling errors by using auxiliary data on population characteristics that are known with a high degree of accuracy.

Sampling weights are also designed to minimize biases arising from differences between respondents and non-respondents by adjusting for nonresponse, and to facilitate the estimation of variances through the use of the replication approach.

Slide 19 of 37

A final weight is required for all sampled persons with a completed background questionnaire and those who could not complete the background questionnaire for literacy-related reasons, but for whom age and gender were collected. Several steps are taken to create the PIAAC final weights.

Slide 20 of 37

First, each sampled household is assigned a household base weight to compensate for differential probabilities of selection. Then, household eligibility and nonresponse rate adjustments are made to reduce potential biases arising from differences between respondents and nonrespondents. These first two steps are employed only for countries that used a screener, like the United States.

Slide 21 of 37

Next, each sampled person is assigned a person base weight to compensate for differential probabilities of selection within the household. Then, for countries that used a registry as a sampling frame, person-level eligibility and nonresponse rate adjustments are made to compensate for sampled persons who refused to participate; were inaccessible; had a physical disability that prohibited their participation; or, did not respond due to literacy-related reasons.

Because of unequal household sizes, sample designs that include the selection of dwelling units have more variability in larger weights compared to directly sampling persons from registries. Therefore, large weights are adjusted by trimming them to a designated cutoff value to reduce their variability, as needed.

Slide 22 of 37

In addition, deficiencies in the sampling frame can result in a sample that fails to cover segments of the target population in proportions representative of those same segments in the population. Therefore, the last step in calculating the final PIAAC weight, once the adjusted household and person level weights have been combined and trimmed, is to compensate for this noncoverage by calibrating person weights to independent control totals. This means that the sampling weights are modified so that the totals on specified characteristics, such as age and gender, agree with corresponding totals known for the population. We will discuss the characteristics used for calibration and nonresponse adjustment next.

Slide 23 of 37

After data collection and data editing, countries conducted analyses to select variables for weighting adjustments that would be most effective in reducing nonresponse bias. Of the countries that provided information, all used age and gender in calibration, as required in the PIAAC Technical Standards and Guidelines. Region was also used in all countries in either calibration or nonresponse adjustment.

In addition, the majority of countries included in their weighting adjustments at least one variable related to education, employment status or nationality, which have been shown to be correlated with proficiency.

Slide 24 of 37

All PIAAC variables selected for benchmarking (or calibration) were required to have reliable control totals available. Countries used data from labor force surveys, censuses, and registries, among other sources, to calibrate PIAAC population totals to accurately represent the total population. The quality of data from these external sources was to have exceeded the quality of data from PIAAC; for example, the standard errors, or more generally, the mean square error of the external estimates needed to be smaller than those of the non-benchmarked estimates from the survey. Control population totals used in the benchmarking process were designed to have the same definition and coverage of the PIAAC target population. More information on the final weights adjustment is available in Chapter 15 of the PIAAC Technical Report, which can be accessed by clicking on the corresponding underlined screen text.

Slide 25 of 37

Statistical analyses are not valid unless the corresponding variance estimators appropriately reflect all of the complex features of the PIAAC sample design, such as stratification and clustering.

In Common Module 4, two standard error calculation procedures were discussed: Replication Techniques and Taylor Series linearization.

To ensure validity, the replication approach is used for estimating variances in analyses of PIAAC data. Replication is a method that calculates appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples, or replicates.

Slide 26 of 37

The choice of the replication method was guided by the particular sample design used in each country. Participating countries have used ONE of four different replication schemes, which include:

- Delete-one jackknife, or JK1;
- Paired jackknife, or JK2;
- Balanced repeated replication, or BRR; and
- Fay's method, which is a variant of the BRR approach.

You can find more information about the number of replications and schemes used by participating countries in Chapter 15 of the PIAAC Technical Report, which can be accessed by clicking on the corresponding underlined screen text.

Slide 27 of 37

In the United States, some 15 variables were used in the final weight adjustment procedures. These included age, gender, region, educational attainment, employment status, country of birth, a measure of linguistic isolation, and the presence of children in the home. Estimates from the American Community Survey were used as benchmarks to calibrate the sample to the total population.

The result of these adjustment procedures is one final weight called SPFWT0.

For variance estimation, the United States used the paired jackknife, or JK2, method with 80 replicate weights.

Accordingly, to calculate standard errors for the United States PIAAC data, you need to select the replicate weights that are associated with the final weight. The replicate weights associated with the final weight SPFWT0 are SPFWT1 through SPFWT80.

More information about the final and replicate weights for the United States is available in the PIAAC 2012 U.S. Main Study Technical Report, which can be accessed by clicking on the underlined screen text, 'U.S. Final Weight.'

Slide 28 of 37

Due to PIAAC's matrix sampling design, individual-level results cannot be provided, because each adult answers only a small number of assessment questions. Therefore, PIAAC provides reliable estimates of proficiency only at the national level, or at the level of large subgroups, such as all females; all employed persons; or, college-educated persons.

Plausible Values, or PVs, is the methodology used to save PIAAC dataset information at the individual case level, which in turn allows the calculation of estimates of proficiency at the national or subgroup level.

It is important to note that each case's PVs reflect not only that particular individual's performance on the items she or he answered; but also, the performance of similar respondents on the rest of the PIAAC assessment.

More information regarding the multiple matrix sampling design can be found in the module titled 'Data Collected Through the PIAAC,' which can be accessed by clicking on the underlined screen text, 'assessment questions.'

Slide 29 of 37

To obtain plausible values, first, each adult's performance on the assessment questions is summarized in the form of a probability curve, or distribution, with a value for every possible score.

The probable score distribution is sometimes bell-shaped, but sometimes it is not. The probable score distribution is based solely on the individual's own performance on PIAAC assessment questions.

Slide 30 of 37

Next, the probability distributions for all adults in a subgroup are used to estimate that group's probability distribution. The estimates for each group's performance are then used to weight each respondent's probability distribution.

Calculations then alternate between weighting each respondent's probable score distribution by the group's performance; and, combining the weighted score locations into new estimates of group performance. These calculations continue until the group performance stops changing. Neither the group performance, nor the weighted respondent performance, is known in advance. They are arrived at simultaneously.

Slide 31 of 37

To avoid having to perform all of these calculations each time an analysis of PIAAC proficiency data is conducted, the results of the iterative process are saved for each respondent in the form of plausible values.

Each of these values is a plausible representation of the performance of an individual with particular background characteristics.

Plausible values are samples from the final weighted probable score distributions for each respondent. As discussed previously, they reflect not only that particular individual's performance on the small number of items she or he answered, but also, the performance of similar respondents on the rest of the PIAAC assessment.

Slide 32 of 37

To review, plausible values, or PVs, are a statistical means to replicate a probable score distribution that summarizes how well each respondent answered a small subset of the assessment items; and, how well other respondents from a similar background performed on the rest of the assessment item pool.

Each individual case in the PIAAC dataset has a randomly chosen set of ten PVs, and all ten PVs must be used together to estimate proficiency. Otherwise, the variability in the predicted outcomes will be understated. The randomly chosen set of PVs best represents the score distribution for a particular subgroup of adults.

For information on the derivation of PVs using item response theory scaling of the cognitive items, or IRT, and the population model used for PIAAC scaling, refer to Chapter 17 of the PIAAC Technical Report, which can be accessed by clicking on the corresponding underlined screen text.

Slide 33 of 37

To summarize, for accurate estimations involving proficiency scores, calculations must account for both the sampling error component, and the variance due to imputation of the proficiency scores.

To account for the sampling error component, you must use the final weight and the corresponding 80 replicate weights. To account for the imputation variance, you must use all ten plausible values.

Slide 34 of 37

One final consideration for analyzing PIAAC data is that missing data can occur when some of the adults selected in the sample are not accessible or refuse to participate, when they fail to respond to a particular survey item, or, because data collected from the sampled adults are contaminated or lost during or after the data collection phase.

All missing data for the PIAAC Background Questionnaire are marked in the dataset as valid skips, don’t know, refused, or not stated/inferred. No Background Questionnaire data in the U.S. national public-use data file or restricted-use data file were imputed. All missing assessment item responses are marked as missing, and no answers were imputed.

All proficiency scores, in the form of Plausible Values, were imputed using item response theory scaling, or IRT, and a population model for each country.

Slide 35 of 37

This table shows the codes for missing values that are used in SAS and SPSS. Especially in SPSS, and when transferring data to other programs, users should be

aware of the missing value codes which include 6, 96, 996, and others shown here, as they may be interpreted as responses, misrepresenting the resulting estimates.

More information about missing cases is presented in the module titled “Considerations for Analysis of PIAAC Data”, which can be accessed by clicking the corresponding underlined screen text.

Slide 36 of 37

This module has provided you with a summary of the PIAAC sample design as it relates to study weights. The module described sampling weights that must be applied to assure that data are representative of the target population, as well as techniques for variance estimation and correctly calculating standard errors for hypothesis testing. The use of scaling in large-scale international assessments was explained, as was the use of plausible values for the analysis of assessment data. Finally, the module described how missing data were handled to ensure accurate analysis.

Slide 37 of 37

Additionally, important resources that have been provided throughout the module are summarized here.

In the next module, you will find more detailed information on special considerations for the analysis of PIAAC data, and explore the functions of the IEA IDB Analyzer. You may now proceed to the next module in the series, or click the “Exit” button to return to the landing page.